Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables

Hervé Déjean, Éric Gaussier

Herve.Dejean@xrce.xerox.com, Eric.Gaussier@xrce.xerox.com

XRCE

6 chemin de Maupertuis

38240 Meylan, France

ABSTRACT. *We present in this article a new method for automatic extraction of bilingual lexicons from comparable corpora. We first anaylze the assumptions underlying the research works led in this domain, and then detail the associated algorithms. Lastly, we evaluate our approach on two different corpora, and show how the combination of our method with standard ones significantly improves the quality of the extracted lexicons.*

KEYWORDS : *bilingual lexicon extraction, comparable corpora, multilingual thesaurus*

RESUME. *Nous proposons dans cet article une nouvelle méthode pour l'extraction de lexiques bilingues de corpus comparables. Pour ce faire, nous revenons tout d'abord sur les hypothèses sous-jacentes aux travaux dans ce domaine, et détaillons ensuite les algorithmes qui en découlent. Enfin, nous évaluons notre approche sur deux corpus aux caractéristiques différentes, et montrons comment la combinaison de notre méthode avec les méthodes standard améliore de façon significative les résultats.*

MOTS-CLES : *extraction de lexique bilingue, corpus comparable, thesaurus multilingue*

Introduction

Si l'extraction de lexique bilingue à partir de corpus parallèle est maintenant un classique en extraction terminologique, et est une activité qui est passée du domaine de la recherche au domaine commercial, l'extraction de lexique bilingue à partir de corpus comparables est plus récente et reste confinée au domaine de la recherche. Les résultats obtenus, s'ils sont encourageants, ne sont pas encore satisfaisants pour une application pratique. [FUNG 98] obtient, pour le couple de langues anglais/chinois, une précision de 76% sur les 20 premiers candidats proposés, score que nous jugeons trop bas pour permettre une révision manuelle. [RAPP 99] atteint 89% sur la paire anglais/allemand en considérant les 10 premiers candidats, mais en évaluant des mots très fréquents de la langue, et en utilisant un corpus de plusieurs dizaines de millions de mots, taille qui peut être difficile à atteindre dans un domaine spécialisé.

Nous proposons dans cet article une nouvelle approche à l'extraction de lexiques bilingues de corpus comparables. En particulier, nous montrons un exemple pratique d'utilisation de thésaurus multilingues pour cette tâche. Nous commençons par revenir sur les hypothèses sous-jacentes aux travaux dans ce domaine, et en formulons une nouvelle qui est à la base de notre méthode. Puis, après avoir détaillé les données que nous utilisons pour nos expériences, nous expliquons notre alternative aux techniques actuelles et la comparons à celles-ci. Enfin, nous montrons que les méthodes sont complémentaires et que la combinaison de ces méthodes permet d'améliorer les résultats.

1. Qu'est-ce qu'un corpus comparable et pourquoi l'utiliser ?

Avant de discuter de l'utilité de corpus bilingues comparables, commençons par en donner une définition dans le cadre de l'extraction de lexique bilingue.

La littérature nous offre plusieurs critères pouvant servir à calculer le degré de comparabilité/comparaison entre deux collections. Nous pouvons utiliser les critères qualitatifs utilisés en stylistique comme le genre, l'auteur, la période, le média [BIBER 93], [SINCLAIR 96], ainsi qu'un grand nombre de mesures quantitatives se basant sur la fréquence des mots [KILGARIFF 01]. Souvent ces critères peuvent s'appliquer aussi bien à des textes monolingues que bilingues. Dans le cadre de l'extraction terminologique bilingue, nous proposons le critère minimal suivant afin d'établir le degré de comparaison entre deux textes *bilingues* :

Deux corpus de deux langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1.

Si ce critère peut aussi être utilisé avec des textes monolingues, en remplaçant la relation de traduction par une relation d'identité, il assure, dans le cas de corpus de langue différente, que l'extraction de lexiques bilingues peut être envisagée, condition suffisante pour notre travail. La notion de sous-partie "non négligeable" est floue, mais elle rend bien compte du continuum qui existe entre des corpus parallèles, des corpus parallèles bruités [VERONIS 00], des corpus comparables à un degré moindre et des corpus non reliés. Notre définition n'est bien sûr pas sans rapport avec les critères traditionnels cités plus haut,

puisque la sous-partie va être d'autant plus grande que les deux corpus portent sur un même domaine et sur une même période : la page internationale de deux quotidiens sur une même semaine, par exemple. Si le décalage dans le temps est grand, plusieurs mois par exemple, le vocabulaire commun sera certainement moindre. On peut considérer que la borne supérieure du degré de comparabilité est obtenue avec un corpus parallèle, puisque la sous-partie est alors maximale, proche de la totalité du vocabulaire.

Mais pourquoi utiliser un corpus comparable plutôt qu'un corpus parallèle ? Le premier des avantages, cité dans la littérature [FUNG 98], est d'ordre pratique : il est plus facile d'accéder à un corpus comparable dans un domaine donné qu'à un corpus parallèle de bonne qualité. Les corpus souvent utilisés dans les travaux précédents sont, comme d'habitude serait-on tenté de dire, les journaux ou les dépêches journalistiques. Nous utilisons dans cet article des corpus dans deux domaines spécifiques : le domaine médical et le domaine des sciences sociales. Les données sont décrites à la section 3.

Un deuxième avantage, plus théorique, est propre au domaine terminologique : les deux parties monolingues d'un corpus comparable ne sont pas liées. Lors d'une traduction, le vocabulaire de la langue source influence le choix du traducteur (les cognats peuvent, par exemple, être privilégiés). L'utilisation de corpus comparable permet d'avoir accès directement à la terminologie monolingue originelle d'un domaine, à l'usage réel des mots dans chaque langue, et évite donc le biais introduit par la traduction.

Nous montrons dans cet article que les résultats obtenus avec un corpus comparable, s'ils ne sont aussi bons que ceux produit avec un corpus parallèle, restent de qualité suffisante pour permettre, pensons-nous, une révision manuelle et donc une utilisation pratique de tels corpus.

2. Sémantique distributionnelle et corpus comparables : retour sur les hypothèses sous-jacentes

Tous les travaux d'extraction de lexiques bilingues de corpus comparables s'inscrivent dans le cadre d'une sémantique distributionnelle, et décrivent le sens d'un mot par sa distribution sur un ensemble de contextes. La mise en relation entre deux mots ou locutions de langues différentes s'établit sur un plan sémantique, et le corpus bilingue est vu ici comme un objet d'acquisition de connaissances et/ou d'adaptation et de mise à jour de ressources lexicales pré-existantes. La facilité de mise en œuvre d'une approche distributionnelle sur corpus, sa robustesse aux données bruitées et partielles extraites des corpus, la possibilité qu'elle offre d'intégrer assez naturellement des ressources existantes, enfin les résultats obtenus par les travaux qui en relèvent, que ce soit pour la recherche documentaire ([RAJMAN 00]) ou à des fins lexicographiques ([GREFENSTETTE 1994, FABRE 98]), en font un des outils les plus populaires de la linguistique de corpus ([HABERT 1997]). Nous voulons revenir ici sur les hypothèses qui sous-tendent l'utilisation de cet outil pour les corpus comparables. Pour cela, nous allons commencer par un retour sur le monolingue.

L'hypothèse à la base des approches distributionnelles d'acquisition de connaissances sémantiques à partir de corpus peut être formulée de la façon suivante :

Hypothèse 1 : *si deux mots ont des distributions similaires, alors ils sont reliés sémantiquement.*[1]

Par distribution, nous entendons ici distribution sur, ou répartition dans un ensemble de contextes. Il est donc nécessaire tout d'abord de définir un ensemble de contextes, à partir desquels la distribution de chaque mot sera calculée, puis définir une mesure de similarité entre distributions. Ce que postule l'hypothèse précédente, c'est que les mots proches au sens de cette dernière mesure de similarité sont reliés sémantiquement. Notons ici que rien n'est dit sur la nature, autre que sémantique, du lien établi. C'est dire que, si l'on s'en tient à cette hypothèse, on ne peut s'attendre à aller plus loin qu'un premier regroupement en sens qu'il faudra ensuite affiner par d'autres moyens.

L'hypothèse **1** telle que nous l'avons formulée correspond à une remarque faite page 24 de [GREFENSTETTE 1994]. Il faut noter que dans cet ouvrage, cette remarque suit l'exposé d'une expérience réalisée dans [LEWIS 1967] qui correspond en fait à une implication inverse entre les termes de l'hypothèse **1** : si deux mots sont reliés sémantiquement, alors leurs contextes, et donc leurs distributions[2] sur ces contextes, sont proches. Ainsi, Lewis *et al*. [LEWIS 1967] montrent que des synonymes, s'ils ne sont pas en relation de cooccurrence directe, sont en relation de cooccurrence avec le même ensemble de mots. Si l'on s'en tient aux relations sémantiques que sont la synonymie, l'antonymie, l'hyper- et l'hyponymie, la contraposée de l'hypothèse **1** n'a rien de surprenant : deux synonymes sont substituables, au sens harrissien du terme, et l'on peut construire des contextes identiques pour ces éléments ; ce qui se dit d'un hyperonyme peut généralement se dire de l'hyponyme, et l'on peut, encore une fois, construire des contextes identiques pour ces éléments ; enfin, deux antonymes peuvent modifier les mêmes éléments, et l'on peut, encore une fois, construire des contextes partagés. Plusieurs remarques s'imposent au sujet des développements précédents : tout d'abord, il existe un écart entre "construire" des contextes partagés, et "trouver en corpus" de tels contextes. Le fait de s'appuyer sur des corpus de grande taille permet de réduire en partie cet écart, mais ne l'abolit pas totalement. De par cet écart, de par les problèmes de polysémie, de délimitation des unités linguistiques et de définition des contextes, les deux hypothèses ci-dessus (**1** et sa contraposée) ne représentent que des tendances, et n'ont valeur de vérité que statistique. Il convient à ce titre de reprendre la formulation de **1** sous la forme suivante :

Hypothèse 1' : *les mots dont les distributions sont les plus similaires à celle d'un mot donné sont, avec une forte probabilité, sémantiquement reliés à ce mot.*

---

[1] Nous trouvons dans [HARRIS 54] la définition suivante de la *distribution* :La *distribution* d'un élément sera définie comme la somme de tous les environnements de cet élément. L'environnement d'un élément A est la disposition effective de ces *co-occurrents*, c'est-à-dire des autres éléments, chacun dans une position déterminée, avec lesquels figure A pour produire un énoncé [page 13].

[2] La distribution sur les contextes peut se calculer de différentes façons, en particulier en terme de présence/absence de cooccurrence avec les éléments du contexte. Ceci explique le "donc" dans notre formulation.

Le seuil de probabilité, ou de rejet d'une hypothèse d'indépendance entre distributions, ne peut être fixé à l'avance, car il dépend d'au moins trois facteurs : le mot que l'on s'est donné, la nature de la relation recherchée, et le corpus d'étude.

La contraposée de **1** et de **1'** est peu intéressante, car non opérationnelle. Même si elle est moins nette que celle de sa contraposée, la validité de **1'** a été établie par différents auteurs, comme nous l'avons déjà signalé. Notons que, en dehors des travaux réalisés explicitement sur l'utilisation de la sémantique distributionnelle en recherche d'information, notamment dans le cadre du modèle DSIR [RAJMAN 00], **1'** se trouve à la base de nombreux autres travaux de recherche documentaire, puisqu'on la retrouve sous-jacente au modèle LSI [DEERWESTER 90] et en partie au modèle GVSM [WONG 85].

Enfin, la dernière remarque qui s'impose concerne la définition du contexte. Comme le notent Habert et al. [HABERT 1997, p. 179], "la taille de la fenêtre [de mots dans laquelle les contextes sont définis] dépend des relations sémantiques que l'on recherche, les cooccurrences à petite, moyenne et grande distance tendant respectivement à faire ressortir des expressions figées ou semi-figées [...], des contraintes de sélection [...] et des mots appartenant au même champ sémantique". L'utilisation de contextes syntaxiques, *i.e.* de contextes définis à partir des mots qui sont en relation syntaxique avec un mot donné, comme l'ensemble des verbes qui admettent pour objet un certain nom, permet de faire apparaître les classes d'objet [GROSS 93] et sous-tend l'apprentissage d'ontologies dans [NEDELLEC 99][3]. Toutefois, même si une telle approche permet de filtrer une partie du bruit associé à l'approche à base de fenêtres et semble bien adapter à des corpus de taille réduite, elle pose problème dès lors que la relation sémantique recherchée permet le passage d'une catégorie grammaticale à une autre. C'est le cas en traduction, où il n'est pas rare d'avoir recours à la transposition [CHUQUET 89], qui consiste, dans sa version de base, à traduire un mot d'une catégorie grammaticale par un mot d'une autre catégorie grammaticale. La transposition est un phénomène assez courant dont on peut voir la source dans les différentes tendances grammaticales entre deux langues et dans les divers processus de formation lexicale. De plus, l'approche syntaxique pure tend à restreindre le champ sémantique utile à la discrimination entre les relations qui peuvent exister entre deux mots. Ainsi, si deux adjectifs antonymes ou opposés modifient les mêmes noms, leurs connotations, prises en compte par des contextes plus larges, restent fort différentes. Il en va ainsi de « chaud » et « froid » dans les phrases suivantes, qui modifient tous deux « mer » mais dont les contextes graphiques, définis au niveau de la phrase, n'ont qu'un élément en commun :

*Ecoutez le vent d'hiver gémir sur la **mer froide** et près de ses bords, ou au dessus des grandes villes, qui, depuis longtemps, ont pris le deuil pour moi.*

[3] Notons à ce sujet que cette approche syntaxique fera également apparaître des classes de noms, support de prédication, partageant un même ensemble de verbes support si l'unité complète (verbe support + nom) n'a pas été identifiée. On voit donc que le problème de la délimitation des unités linguistiques est primordial dans ce type d'approches. Doit-il pré-exister ou co-exister avec celui de l'acquisition des relations sémantiques ? Cette question nous semble entière à l'heure actuelle.

*Puis, lentement, la **mer chaude** des Caraïbes m'accueille, m'enlace, m'entraîne au fond des eaux, au royaume des coraux*[4].

La remarque précédente vaut également dès lors que l'on se place dans une perspective bilingue, pour les contextes graphiques de taille réduite, qui, s'ils permettent le repérage d'expressions figées, apportent peu d'éléments pour l'identification de la traduction de ces unités. De par ces raisons, et de par le manque de ressources, et en particulier d'analyseurs syntaxiques, disponibles dans diverses langues, les travaux en acquisition de lexiques bilingues à partir de corpus comparables se reposent sur des contextes graphiques de taille moyenne, en général plusieurs phrases, inférieure toutefois au paragraphe et au document, les deux unités standard de la recherche documentaire. Signalons néanmoins le travail réalisé dans [SHAHZAD 99] dans lequel les auteurs cherchent à localiser, mais pour des raisons de performance informatique, une région de petite taille dans laquelle se trouve la traduction recherchée.

Il nous semble qu'une recherche devrait être menée, à la fois en monolingue et en bilingue, sur le lien entre les relations sémantiques, les types de contexte à considérer (syntaxique et/ou graphique) et les mesures de similarité à même de tenir compte d'attributs différents. Ainsi, la synonymie implique une similarité de contextes sur tous les types d'attributs, alors que l'antonymie ne semble l'impliquer que pour les attributs syntaxiques. Une ébauche d'une telle réflexion se trouve dans [HABERT 97], mais aucune recherche ne s'est, à notre connaissance, concentrée sur ce point. Une telle recherche dépasse également le cadre du présent article, et nous voulons présenter ici comment l'hypothèse **1'** se traduit dans un cadre bilingue.

Lorsque la relation sémantique que l'on cherche entre unités correspond à une relation de traduction se pose le problème du passage d'une langue à une autre. Si, dans le cas des corpus parallèles, l'espace de recherche des unités à mettre en relation se réduit à la phrase ou à un ensemble de phrases, aucune contrainte d'alignement ne vient, dans le cas du corpus comparable, réduire l'espace de recherche que constitue le corpus dans son ensemble. C'est dire que n'importe quelle unité d'une langue peut être mise en relation, *a priori*, avec n'importe quelle unité d'une autre langue. Le passage d'une langue à une autre dans le cas des corpus comparables se fait par l'intermédiaire de ressources bilingues existantes, dictionnaires ou thésaurus, ressources souvent partielles et/ou peu adaptées au corpus. La méthode est la suivante : on commence par extraire l'ensemble des contextes de chaque mot, dans chacune des langues ; une fois cet ensemble extrait, on traduit chaque élément de chaque contexte de chaque mot (on parlera, dans la suite, de **vecteur de contexte** pour désigner, pour un mot donné, l'ensemble des éléments de son contexte) dans une des deux langues ; on effectue enfin la comparaison entre mots sur la base de leurs vecteurs de contextes, originaux ou traduits, comme dans le cas du monolingue. L'hypothèse sous-jacente à cette approche peut se formuler de la façon suivante :

[4] Ces deux phrases sont respectivement tirées, et adaptées, des « Chants de Maldoror », de Lautréamont, pour la première, et de « L'Accident », une nouvelle de M. Warner-Vieyra, pour la seconde.

Hypothèse 2 : *les mots de la langue l_1 dont les distributions normalisées sont les plus similaires à la distribution d'un mot donné de la langue l_2 sont, avec une forte probabilité, traduction de ce mot.*

Par « distributions normalisées », nous entendons ici distributions sur des vecteurs de contexte traduits. On peut voir qu'un des inconvénients de cette approche est que si aucun des éléments du vecteur de contexte d'un mot ne peut être traduit avec les ressources disponibles, alors on ne pourra trouver de traduction à ce mot. De plus, aucune adaptation des ressources au corpus n'est réalisée, les ressources étant utilisées de la même façon, avec la même information, quel que soit le corpus. On voit donc que cette approche répond mal aux problèmes récurrents d'inadéquation de ressources existantes à un corpus donné.

Si l'on considère une entrée d'un dictionnaire ou d'un thésaurus bilingue comme une pièce à deux faces, l'une dans une langue et l'autre dans l'autre langue, alors il est possible de construire, pour chaque entrée, un vecteur de contexte à double face, formé respectivement du vecteur de contexte, en langue l_1, du mot de l_1 de l'entrée, et du vecteur de contexte, en langue l_2, du mot de l_2 de l'entrée. Les similarités entre mots de *l1* et *l2* peuvent alors être calculées sur la base de la similarité de leurs vecteurs de contexte avec une ou plusieurs entrées de la ressource à disposition, la comparaison avec l'entrée se faisant à partir du vecteur de contexte de l'entrée dans la même langue.

L'hypothèse qui sous-tend une telle approche peut se formuler de la façon suivante :

Hypothèse 3 : *deux mots de l_1 et l_2 sont, avec une forte probabilité, traduction l'un de l'autre si leurs similarités avec les entrées des ressources bilingues disponibles sont proches.*

Les désavantages que nous avons mentionnés ci-dessus ne subsistent pas avec cette approche. On étudie bel et bien le comportement des entrées des ressources sur le corpus, à partir de la construction de leur vecteur de contexte. Les ressources utilisées sont donc adaptées au corpus en considération. De plus, le fait qu'aucun mot d'un vecteur de contexte donné ne puisse être traduit avec les ressources à disposition n'empêche pas de trouver une traduction pour le mot lui-même, par l'intermédiaire des similarités avec les vecteurs de contexte des entrées des ressources.

Nous présentons plus loin (section 4) une méthode fondée directement sur l'hypothèse 3 et la comparons aux méthodes standard, fondées sur l'hypothèse 2 en section 5. Nous allons maintenant décrire les données que nous avons utilisées.

3. Description des donnés utilisées

3.1 Les données

Pour nos expériences, nous avons utilisé deux collections. La première nous a été fournie dans le cadre du projet MUCHMORE (http://muchmore/dfki.de). Ce projet porte sur la recherche d'information multilingue dans le domaine médical. Nous avons utilisé un corpus composé de 845 résumés d'articles scientifiques provenant de la base de données MEDLINE (http://www4.ncbi.nlm.nih.gov/PubMed) rédigés en anglais et en allemand et

portant sur le domaine de la chirurgie, ce qui correspond environ à 100 000 mots pour chaque langue. Les résumés anglais sont en général des traductions des résumés allemands. Dans de nombreux cas toutefois, la version anglaise est une reformulation complète du résumé allemand, ce qui rend difficile un alignement au niveau des phrases. Nous considérons ce corpus comme étant un corpus parallèle bruité, selon la terminologie de [FUNG 98]. Le tableau 1 montre un exemple de traduction de résumé.

Résumé allemand	**Résumé anglais**
Zusammenfassung .	
Anhand einer Kasuistik wird eine Herzbeuteltamponade bei Anlage eines zentralen Venenkatheters via V. subclavia beschrieben .	A case report demonstrates the complication of pericardial tamponade during the installation of a central venous catheter via the subclavian vein .
Zur Senkung der hohen Letalität dieser seltenen Komplikation werden schnell durchzuführende diagnostische Maßnahmen sowie die adäquate Therapie der Perikarddrainage angegeben .	To reduce the high mortality of this rare complication , quickly applicable diagnostic measures and adequate therapy of pericardiocentesis are indicated .
Nur die Kenntnis dieser schwerwiegenden Komplikation als mögliche Ursache einer akuten Verschlechterung des Allgemeinzustands des Patienten bei Anlage eines zentralen Venenkatheters entscheidet über das Überleben des Patienten .	Prompt recognition and treatment of pericardial tamponade are imperative if a disastrous outcome is to be prevented .

Tableau 1 : Un résumé allemand et sa traduction anglaise.

Si les deux premières phrases fournissent deux bons alignements pour l'extraction de lexique bilingue, l'auteur a totalement reformulé la troisième phrase, qui offre peu d'intérêt pour l'extraction de lexique bilingue. On peut aussi noter que, suivant notre définition d'un corpus comparable, la dernière phrase ne constitue pas un corpus comparable, puisque aucun mot allemand ne trouve son équivalent dans la partie anglaise. Au niveau du résumé toutefois, le degré de comparabilité reste fort, une grande partie des termes allemands se retrouvant dans la partie anglaise.

Associés à ce corpus, nous utilisons comme ressource bilingue le dictionnaire ELRA (http://www/icp.grenet.de/ELRA/home.html) anglais-allemand ainsi que le thésaurus anglais du domaine médical MeSH : Medical Subject Headings, (http://www.nlm.nih.gov/mesh/MBrowser.html), et sa version allemande, DMD, fournie par le Deutsches Institut fuer Medezinische Dokumentation und Information (http://www.dimdi.de). Nous avons extrait ces deux thésaurus (version 2000) du méta thésaurus UMLS (Universal Medical Language System : http://www.nlm.nih.gov/mesh/meshhome.html). À travers ce méta thésaurus, les entrées de MeSH et DMD sont alignées. Nous avons donc à notre disposition un thésaurus bilingue

dans le domaine médical. Le nombre d'entrées alignées est d'environ 15000 (MeSH comprend en fait plus de 200000 entrées, mais le nombre d'entrées dans DMD est beaucoup plus faible).

Comme deuxième jeu de données, nous avons utilisé les données fournies par la tâche GIRT de CLEF 2002 (http://clef.iei.pi.cnr.it:2002/). Nous dirons simplement ici que cette tâche consiste à retrouver un ensemble de documents allemands pour une liste de requêtes anglaises (recherche d'information multilingue). Ici aussi nous utilisons le dictionnaire anglais-allemand ELRA comme ressource bilingue générale, et le thésaurus GIRT mis à disposition des participants à CLEF (environ 10000 entrées allemandes avec leur traduction anglaise). Le thésaurus et les corpus portent sur les sciences sociales. Le corpus est constitué d'une collection d'articles scientifiques en allemand (86000 articles). Les titres de ces articles sont traduits en anglais, ainsi que 6% des corps des articles. Les titres constituent donc un corpus parallèle. Afin de constituer notre corpus comparable, nous avons sélectionné dans le BNC (British National Corpus), un sous-ensemble de taille équivalente au corpus allemand, et contenant tous les mots des requêtes anglaises utilisées pour les évaluations 2000 et 2001. En effet, les traductions de ces mots, utilisées dans les requêtes allemandes, apparaissent dans le corpus allemand, ce qui garantit bien un corpus comparable.

Ces deux jeux de données, que nous appellerons MUCHMORE et GIRT, sont de tailles et de types différents : 100000 mots d'un côté, avec un degré de comparabilité élevé, 8 millions de l'autre, avec un degré de comparabilité plus faible. Ces tailles et types différents vont nous permettre de mieux évaluer le comportement des méthodes développées. À titre de comparaison, [RAPP 99] utilisait un corpus de 85 millions de mots (quotidiens allemands et anglais), [FUNG 9X] deux ans du Wall Street Journal et du Nikkei Financial News, quotidien japonais.

3.2 Le prétraitement linguistique

3.2.1 Le prétraitement morphologique.

Chaque corpus a été prétraité de façon à normaliser les mots qu'il contient. Premièrement, les corpus sont étiquetés et lemmatisés afin d'extraire les lemmes des mots pleins (noms, verbes, adjectifs, adverbes). Seuls ces mots sont pris en compte par nos algorithmes d'extraction. De plus, dans la mesure où nous nous sommes concentrés, pour ce travail, sur des mots simples et avons laissé de côté les termes comprenant plusieurs mots, nous avons segmenté les composés allemands. Nous présentons ici l'approche suivie pour ce faire.

3.2.2 La segmentation des composés allemands

Une première étape dans la segmentation des composés allemands repose sur l'utilisation d'un lexique de composés dans lequel les segmentations ont été établies par un lexicographe. Toutefois, cette étape laisse une certaine proportion, environ 25%, de mots composés allemands non segmenté, comme par exemple le mot *Adhaesionsileubehandlung*. Pour améliorer cette décomposition, nous avons introduit deux heuristiques. La première, de nature morphologique, est fondée sur la règle suivante :

w = XABY → XA_Y, avec A=[ung|heit|keit|schatf|aet|ion|aeten|osen], et B=[s|]

qui s'interprète de la façon suivante : un mot *w*, de forme XABY se décompose en deux parties, XA et Y (nous avons adopté ici la syntaxe standard des expressions régulières). La deuxième heuristique permettant de segmenter est fondée sur la présence de sous chaînes d'un mot dans le corpus : si les sous-chaînes obtenues par segmentation d'un mot *w* apparaissent en tant que mots dans le corpus, alors le mot est décomposé en ses sous-chaînes. Par exemple, le mot *Adhaesionsileusbehandlung* est segmenté en *Adhaesionsileus* | *behandlung*, ces deux mots apparaissant dans notre corpus. La segmentation est réappliquée sur chaque sous-chaîne récursivement. Ainsi *Adhaesionsileus* est à son tour segmenté en *Adhaesion ileus* grâce cette fois à la première heuristique (présence de la chaîne *ions*). Seuls les mots de plus de 7 caractères et sous-chaînes de plus de 3 caractères sont pris en compte par cette heuristique, ceci afin d'éviter une sursegmentation.

Avec l'application de ces deux heuristiques, le nombre de types allemands passe de 14700 à 10500. Une estimation montre que 1000 à 2000 types restent à segmenter.

3.3 La méthode d'évaluation

Pour évaluer nos algorithmes, nous avons construit un lexique de référence extrait manuellement à partir du corpus. Pour notre première collection (domaine médical), un lexique de 1800 mots a été utilisé, et pour notre deuxième collection (sciences sociales), nous avons utilisé les 180 mots des requêtes 200 et 2001 de la tâche GIRT. La mesure retenue consiste à prendre une liste de *n* candidats fournis par une méthode, et à y cherche la traduction donnée dans le lexique de référence. La valeur de n varie de 1 à 20, cette gamme de valeurs fournissant des listes assez courtes pour permettre à un lexicographe de valider la bonne traduction, si elle se trouve dans la liste.

Les mesures standard de précision, rappel et F-mesure (F-1 en l'occurrence) sont utilisées pour comparer les mesures entre elles.

4. L'extraction de lexique bilingue

Dans cette section, nous décrivons les deux méthodes associées aux hypothèses 2 et 3 que nous avons formulées plus haut. Comme nous l'avons déjà mentionné, l'hypothèse 2 est à la base de la majorité des travaux existants en extraction de lexiques bilingues de corpus comparables, et nous appellerons la méthode associée méthode par traduction directe, en opposition à la méthode par similarité interlangue, à la base de notre approche et associée à l'hypothèse 3.

4.1 L'approche par traduction directe

Comme nous l'avons déjà souligné, c'est l'approche suivie par la majorité des travaux dans le domaine ([PETERS 95], [TANAKA 1996], [SHAHZAD 99], [RAPP 99] et [FUNG 98, 00]). Notre implémentation de l'algorithme général mentionné en section 2 passe par les étapes suivantes :

1. pour chaque mot *w*, nous construisons son vecteur de contextes en considérant tous les mots pleins qui occurrent avec *w* dans une fenêtre de *n* phrases autour de la phrase contenant *w*. Chaque mot *i* du contexte de *w* est ensuite pondéré par une mesure de son association avec *w*. Dans nos expériences, *n* est pris égal à 1 (on considère donc la phrase courante, la phrase précédente et la suivante), et la mesure d'association retenue est le test du rapport de vraisemblance [DUNNING 93], qui correspond ici à une information mutuelle généralisée, et que nous noterons v_{wi} ;
2. les vecteurs de contexte des mots de l_1 sont ensuite traduits en langue l_2 à l'aide des ressources à disposition, décrites ci-dessus, en laissant les poids de l'étape 1 inchangés. Quand plusieurs traductions sont possibles, nous les considérons toutes avec le même poids (nous revenons plus bas sur ce choix) ;
3. les vecteurs de contexte (traduits et originaux) sont ensuite comparés sur la base d'une mesure de similarité. Dans nos expériences, cette mesure de similarité est le cosinus de l'angle formé par les deux vecteurs sur l'espace vectoriel défini par l'ensemble des éléments des vecteurs de contexte de l_2. Cette mesure est définie par :

$$sim(w_1, w_2) = \frac{\sum_i v_{w1i} v_{w2i}}{\sqrt{\sum {v_{w1i}}^2 {v_{w2i}}^2}}$$

4. les valeurs des similarités ainsi obtenues sont normalisées afin de fournir un lexique de traduction probabiliste, $P(w_2|w_1)$ (où w_1 et w_2 sont des mots quelconques des corpus de l_1 et l_2 respectivement).

Un certain nombre de remarques s'imposent ici : tout d'abord, le choix des mesures, rapport de vraisemblance et cosinus, a été fait sur la base du bon comportement de ces mesures dans des travaux précédents. D'autres choix auraient pu être faits, mais notre but ici est de comparer deux méthodes générales, et pas un ensemble de mesures. La traduction des vecteurs de contexte proposée à l'étape 2 peut paraître simple. En pratique, elle n'est pas si aisée, comme le note déjà [FUNG 98], et dépend de la nature des entrées de la ressource bilingue utilisée. Prenons par exemple le dictionnaire anglais/allemand ELRA. Chaque entrée de ce dictionnaire est identifiée par un numéro unique. Une entrée dans une langue peut comporter plusieurs termes qui peuvent être à leur tour simple (un seul mot) ou multi-mots. La méthode la plus simple consiste à traduire un mot par l'ensemble de ses traductions, et affecter comme valeur à chaque traduction le poids du mot traduit. Ce qui introduit donc du bruit. Pour essayer de résoudre ce problème de traduction, [FUNG 98] introduit un score de confiance entre chaque pair de traduction fonction de l'ordre de traduction offert par le dictionnaire, qui modifie la fonction de similarité entre deux vecteurs. Cette modification suppose que la première traduction trouvée dans le dictionnaire a plus de poids que la seconde, hypothèse non vérifiée avec nos ressources. Le tableau 2 montre le vecteur de contexte original du mot *Leber*, sa traduction, et le vecteur de contexte de *liver*, traduction de *Leber* (ces vecteurs sont restreints aux 10 premiers

éléments). Pour des raisons de lisibilité, les poids ne sont pas donnés). Si l'on voit que certains mots sont correctement traduits (*Metastate, Hepatitis*), certains, ne se trouvant pas dans le dictionnaire, disparaissent du vecteur du vecteur traduit (*Resektion*). Le mot *Transplantation* est traduit par *transplant*, et non pas *transplantation*, utilisé dans le corpus.

vecteur de Leber (10 premiers)	Transplantation, Resektion, Metastase, Visualisierung, Hepatitis, Computer, Arterie, Shunt, Planung, metabolisch
vecteur précédent traduit (10 premiers)	Transplant, secondary, metastasis, program, hepatitis, computer, artery, design, scheme, planning
vecteur de liver (10 premiers)	transplantation, resection, metastasis, survival, visualization, hepatitis, hepatic, orthotopic, dimensional, shunt

Tableau 2 : vecteurs de contextes du mot *Leber*

Un autre problème se pose aussi lors de la traduction: comment gérer les entrées multi-mots. Puisque les vecteurs de contextes sont constitués de mots simples, les entrées multi-mots du dictionnaire bilingue ne peuvent pas être exploitées directement. Nous avons testé plusieurs méthodes de traduction, de la prise en compte exclusive des mots simples au produit cartésien entre les différentes parties des termes. Les meilleurs résultats ont été obtenus avec la prise en compte des mots simples seulement, et c'est ce choix que nous faisons pour la suite ([RAPP 99] a recours à un choix similaire). Cela signifie donc une exploitation partielle des ressources, comme le montrent les distributions données dans le tableau 3. Une alternative, non testée pour l'instant, serait de prendre en compte les entrées multi-mots des ressources bilingues dans la construction des vecteurs de contexte en intégrant les multi-mots comme nouvelles dimensions de l'espace. Mais se pose alors le problème de leur reconnaissance.

	# total de termes	# termes simples
ELRA (angl.)	47485	30935 (65%)
ELRA (all.)	46194	39932 (87%)
Thes. GIRT (angl.)	9706	2693 (27%)
Thes. GIRT (all.)	16038	13562 (84%)

Tableau 3 : évaluation mots simples/multiples dans le dictionnaire ELRA et le thésaurus GIRT.

4.2 L'approche par similarité interlangue

Comme nous l'avons signalé plus haut, on peut envisager d'autres façons de relier des mots de langues différentes à partir d'un corpus comparable et d'une ressource bilingue. L'hypothèse 3 indique une telle possibilité, et nous allons maintenant présenter un modèle fondé sur cette hypothèse.

4.2.1 Modèle général

De manière générale, une ressource bilingue établit un pont entre plusieurs langues à travers les correspondances entre mots ou classes conceptuelles qu'elle contient (dans le cas d'un thésaurus, une classe conceptuelle regroupe l'ensemble des synonymes et des points de vue sur un concept donné). Par exemple, la classe C0751521 de MeSH, pour laquelle l'entrée principale est *splenic neoplasms*, contient aussi *cancer of spleen, splenic cancer* et *spleen neoplasms*). La correspondance entre les entrées à travers plusieurs langues peut être un-vers-un, à une entrée dans une langue correspond une et une seule entrée dans l'autre langue, ou n-vers-m, une entrée dans une langue peut être associée à plusieurs entrées dans l'autre langue. En désignant par w_i (respectivement C_i) un mot (respectivement une entrée de la ressource bilingue) en langue i, et en notant $P(w_2|w_1)$ la probabilité d'associer le mot w_2 au mot w_1, nous avons :

$$\begin{aligned} P(w_2 \mid w_1) &= \sum_{C_1} P(C_1, w_2 \mid w_1) \\ &= \sum_{C_1} P(C_1 \mid w_1) P(w_2 \mid C_1, w_1) \\ &= \sum_{C_1} \sum_{C_2} P(C_1 \mid w_1) P(C_2, w_2 \mid C_1, w_1) \\ &= \sum_{C_1} \sum_{C_2} P(C_1 \mid w_1) P(C_2 \mid C_1, w_1) P(w_2 \mid C_2, C_1, w_1) \\ &= \sum_{C_1} \sum_{C_2} P(C_1 \mid w_1) P(C_2 \mid C_1) P(w_2 \mid C_2, C_1, w_1) \end{aligned} \qquad (1)$$

La dernière étape dans la dérivation ci-dessus utilise le fait que la correspondance entre entrées soit donnée, et dépend de ces entrées seulement. La forme finale de l'équation (1) s'interprète de la façon suivante : à partir du mot w_1, on sélectionne une entrée C_1 dans notre ressource bilingue sur la base de la distribution $P(C_1|w_1)$; on sélectionne ensuite les entrées en langue 2 correspondantes à partir de $P(C_2|C_1)$; enfin, on choisit les mots w_2 sur la base de $P(w_2|C_2,C_1,w_1)$.

Dans le cas où la correspondance entre entrées est un-vers-un, comme c'est le cas avec nos ressources, alors la distinction entre C_1 et C_2 n'a plus lieu d'être, et l'équation précédente se simplifie et fournit :

$$P(w_2 \mid w_1) = \sum_{C} P(C \mid w_1) P(w_2 \mid C, w_1) \qquad (2)$$

La dépendance sur w_1 dans la dernière distribution de probabilité ($P(w_2|C,w_1)$) permet de privilégier, au sein d'une même entrée, une lexicalisation particulière. On pourrait l'utiliser, par exemple, pour choisir *spleen neoplasms*, plutôt que *cancer of spleen*, comme traduction de *Milztumoren* depuis la classe C0751521. Toutefois, une telle distinction dépasse le cadre du présent article, et nous faisons ici l'hypothèse que w_2 est indépendant de w1, une fois l'entrée *C* choisie, ce qui donne :

$$P(w_2 \mid w_1) = \sum_C P(C \mid w_1) P(w_2 \mid C, w_1) \qquad (3)$$

qui est l'équation que nous utilisons par la suite.

4.2.2 Estimation des paramètres

L'estimation des paramètres de l'équation (3), à savoir $P(C|w_1)$ et $P(w_2|C)$, passe par le calcul de la similarité entre les vecteurs de contexte des mots w_1 et w_2, et ceux des entrées de notre ressource bilingue, comme le suggère l'hypothèse 3 (et même l'hypothèse 1' puisque nous sommes dans un cadre monolingue pour ces similarités). Les vecteurs de contexte des mots du corpus sont calculés de la même manière que pour la méthode par traduction directe. Les similarités entre vecteurs de contexte sont encore une fois calculées sur la base du cosinus, puis normalisées de façon à fournir des distributions de probabilité. Si l'entrée de langue li de notre ressource ne contient qu'un mot, son vecteur de contexte est celui de ce mot. Si l'entrée contient un terme composé, comme *liver disease*, nous construisons un vecteur de contexte comme la conjonction des vecteurs de contexte des mots simples, en normalisant les poids par le nombre de mots de l'unité (deux ici). Ainsi, le vecteur de contexte de *liver disease* contiendra seulement les mots qui apparaissent à la fois dans les vecteurs de contexte de *liver* et *disease*, dans la mesure où l'unité dans son ensemble est un concept plus étroit que ses constituants. Enfin, si l'entrée contient plusieurs mots simples ou plusieurs termes composés, nous considérons le vecteur de contextes obtenu par la disjonction des vecteurs de contexte de chacun des mots simples et termes composés, en normalisant les poids par le nombre de mots simples et termes composés de l'entrée.

L'exemple suivant illustre un cas intéressant : le mot allemand *Aktinomykose* apparaît parfois dans notre corpus sous la forme *Actinomykose*, alors que seule la première forme est présente dans notre ressource (entrée C0001261). Toutefois, par l'utilisation des vecteurs de contexte, nous retrouvons C0001261 comme classe la plus proche (au sens de $P(C|w)$) de *Actinomykose* (et également de sa traduction, *Actinomycosis*). On voit donc que l'estimation des paramètres repose directement sur les vecteurs de contexte extraits en 4.1.

4.2.3 Recherche directe et recherche structurelle

L'équation (3) suggère de calculer l'association entre deux mots w_1 et w_2 à partir de l'ensemble des entrées de notre ressource. Nous présentons plus loin une évaluation des lexiques obtenus en faisant varier le nombre d'entrées retenues, à partir de leur proximité

(au sens de *P(C|w)*) avec w_1. Toutefois, si une relation entre un mot et une entrée n'est pas pertinente, cette méthode a le désavantage d'introduire du bruit dans le calcul d'association. De plus, elle ne tient pas compte des dépendances entre entrées de la ressource bilingue, dépendances structurelles fortes dans le cas des thésaurus. Nous présentons ici une méthode pour prendre en compte de telles dépendances.

De manière intuitive, si deux classes[5] sont à la fois proches d'un mot donné w1 et proches entre elles dans le thésaurus, alors la traduction de w1 a de grandes chances d'être associée à ces classes et aux classes qui les relient dans le thésaurus. Par exemple, si w1 sélectionne les deux classes *Hepatitis* et *Cirrhosis*, alors il y a de grandes chances que la traduction relève de *Liver Diseases*, la classe père des deux classes précédentes. Nous utilisons cette intuition de la manière suivante :

1. à partir d'un mot donné w_1, nous sélectionnons les *n* classes les plus proches, au sens de *P(C|w)*. Ceci nous fournit un premier ensemble de classes E_0 ;
2. pour chaque paire de classes de E_0, nous ajoutons à la liste *E* des classes retenues, originellement vide, toutes les classes se trouvant sur le chemin minimal reliant les deux classes. Le chemin minimal est défini à partir de la structure du thésaurus et comprend les deux classe de départ (notons que le thésaurus est en général une forêt d'arbres, et que la définition du chemin minimal ne pose donc pas de problème) ;
3. l'association entre mots est alors définie par :

$$P(w_2 \mid w_1) = \sum_{C \in E} P(C \mid w_1) P(w_2 \mid C, w_1) \qquad (4)$$

Nous n'utilisons pas ici de véritable mesure de distance entre classes d'un thésaurus (comme il est fait par exemple dans [MAYNARD 98]), mais nous reposons directement sur la structure du thésaurus. Le résultat de cet algorithme est de fournir un ensemble de sous-arbres dérivés des 15 sous-thésaurus de MeSH, dont les libellés sont donnés figure 1. Pour une liste initiale de vingt classes (*n=20*), nous obtenons en moyenne 4 sous-arbres par mot, et le nombre final moyen de classes par mot est d'environ 30.

Pour le mot *Leber*, (CUI C0023884 dans UMLS), cet algorithme introduit 13 nouvelles classes si la liste initiale est composée de 20 classes formant 4 sous arbres. Le tableau 4 montre la différence de qualité entre la méthode qui n'utilise pas de structure, que nous appellerons « plate », et la méthode qui repose sur la structure du thésaurus, que nous appellerons « hiérarchique ». Cette dernière offre deux candidats corrects (liver, hepatic) dans les 5 premiers candidats, alors que pour la méthode « plate » il faut aller jusqu'aux rangs 21 et 87 pour les trouver.

[5] Nous utilisons classe dans le sens d'entrée d'un thésaurus. La discussion qui suit concerne plus particulièrement cette ressource, les dictionnaires n'ayant en général d'autres structures que celle induite par l'ordre alphabétique.

5. Évaluation

Nous évaluons dans cette section les différentes approches décrites précédemment. Les premiers résultats concernent le corpus médical, puis nous présenterons les résultats avec le corpus GIRT.

Anatomy [A] , Organisms [B], Diseases [C], Chemicals and Drugs [D],
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E],
Psychiatry and Psychology [F], Biological Sciences [G], Physical Sciences [H],
Anthropology, Education, Sociology and Social Phenomena [I],
Technology and Food and Beverages [J], Humanities [K], Information Science [L],
Persons [M], Health Care [N], Geographic Locations [Z]

Figure 1 : les 15 sous thésaurus de MeSH

PLAT200	HIER20
hepatocyte	liver
neoplam	orthotopic
enormous	hepatic
hepatic (rang : 27)	survival
liver (rang 81)	metastasis

Tableau 4 : Comparaison entre la méthode HIER20 et PLAT200 pour le mot *Leber* (foie).

5.1 Corpus médical

La ressource bilingue utilisée est le thésaurus MeSH. Commençons par une évaluation de la méthode standard. Utilisant MeSH, un tiers des mots des vecteurs de contextes sont traduits, souvent les composantes majeures du vecteur. Les résultats sont montrés tableau 5. Si ces résultats sont en deçà des évaluations fournis dans les articles d'écrivant cette méthode([FUNG 00] 76%, [RAPP 98] 89%), l'explication provient sans doute de la différence de taille entre notre corpus médical et les corpus utilisés précédemment. Nous renvoyons le lecteur à la section suivante qui décrit l'évaluation utilisant le corpus GIRT (8 millions de mots), qui compare nos résultats avec la méthode standard.

Nombre de candidats	score
10 meilleurs	44
20 meilleurs	57

Tableau 5 : Corpus médical : évaluation de la méthode standard (traduction des vecteurs).

Évaluons maintenant la première version de notre méthode, appelée plate par opposition à la méthode utilisant la hiérarchie. Le paramètre de la méthode étant le nombre d'entrées, le tableau 6 montre les différentes évaluations obtenues pour des valeurs allant de 1 à 200. La stabilité des résultats est obtenue à 50 entrées. Si le nombre de classes est trop faible (1, 5), les résultats ne sont pas optimaux. La prise en considération de nouvelles entrées après 50 ne modifie pas les résultats. Si l'on compare ces résultats avec la méthode standard, celle-ci est légèrement en faveur de cette dernière méthode. Particulièrement pour l'évaluation sur les vingt meilleurs candidats.

	1	5	10	50	100	150	200
10 meilleurs	33	38	39	**43**	43	43	43
20 meilleurs	39	45	46	**51**	51	51	51

Tableau 6 : Corpus médical : évaluation en fonction du nombre d'entrées considérées : méthode plate.

	HIER5	HIER10	HIER20	HIER50	HIER100
10 meilleurs	39	43	47	**50**	47
20 meilleurs	43	51	59	**63**	59

Tableau 7 : évolution des résultats en fonction du nombre d'entrées : méthode hiérarchique.

Si nous regardons maintenant les résultats obtenus avec la méthode hiérarchique (tableau 7), nous voyons que ceux-ci sont meilleurs que la méthode standard, mais les résultats sont toujours très en deçà des résultats fournis par un corpus parallèle.

Si nous comparons les trois méthodes (standard, plate, hiérarchique), nous pouvons remarquer que

1. pour un nombre d'entrées équivalent, la méthode hiérarchique offre de bien meilleurs résultats que la méthode plate. De plus, la méthode hiérarchique avec 20 ou 50 entrées fournit de meilleurs résultats que la méthode plate quel que

soit le nombre d'entrées choisi pour cette dernière. Il semble donc que l'utilisation de la hiérarchie comme moyen de sélection d'entrées soit très efficace.

2. la solution optimale de la méthode hiérarchique (HIER50 :50/63) offre de meilleurs résultats que la méthode standard (44/57). Plusieurs points peuvent expliquer ces meilleurs résultats. Premièrement, l'utilisation des ressources lexicales bilingue comme intermédiaire entre les mots source et les mots cibles peut jouer le rôle de filtre, éliminant une certaine quantité du bruit inhérent à la méthode basée sur les vecteurs de contexte. Deuxièmement, notre méthode exploite mieux les ressources multilingues disponibles, en particulier les entrées multi-mots qui ne sont pas exploitées dans la méthode standard. Nous reviendrons sur ce point dans la section suivante.

5.2 Corpus GIRT.

Pour ce corpus, les ressources bilingues utilisées sont le dictionnaire ELRA et le thésaurus GIRT. Dans cette expérience, nous comparons la méthode classique avec notre méthode plate. La différence entre les deux méthodes est très significative, alors qu'elle ne l'était pas pour le corpus médical. Une explication possible est la différence de taille entre les deux corpus (100000 et 9 millions de mots), un corpus de grande taille permettant une construction des vecteurs de contextes plus représentative. Les 180 mots pris comme évaluation ont de fait des fréquences assez élevées (très souvent supérieures à 100 voire 1000). L'évaluation avec le corpus médical prend en compte un nombre non marginal de mots de faible fréquence ainsi que des hapax. Cette différence peut aussi s'expliquer par la forte présente de multi-mots dans le thésaurus. Le tableau 3, présentée section 4, montre que les multi-mots représentent plus des trois-quarts des entrées anglaises. Comme seuls les mots simples sont utilisés dans la traduction, le thésaurus n'est alors que peu exploité avec l'approche standard. Par contre notre approche permet une exploitation de ces multi-mots, puisque nous calculons aussi la similarité entre un mot et une entrée comportant des multi-mots. Notre méthode nous permet donc une meilleure exploitation des ressources bilingues disponible.

	Méthode plate	**Méthode standard**
10	79	35
20	84	42

Tableau 8 : Évaluation méthode plate/standard pour le corpus GIRT.

6. Et en combinant les approches ?

Au lieu de considérer notre approche et la méthode standard, nous pouvons essayer de voir si elles sont complémentaires : en combinant les deux méthodes, obtient-on de meilleurs résultats que chacune des méthodes prises isolément ? La combinaison utilisée est une simple combinaison linéaire entre les ressources. Des expériences sont en cours pour

permettre une optimisation de ces ressources en fonction de différents paramètres comme la fréquence du mot dans le corpus, ou sa proximité avec le thésaurus (par exemple, si la proximité avec le thésaurus est faible, alors la méthode standard devrait être privilégiée).

Le tableau 9 nous montre que cette combinaison est très fructueuse, puisqu'un gain de très de 20% en absolu est obtenu sur les meilleurs résultats des méthodes prises individuellement. De plus, une combinaison de la méthode standard avec n'importe laquelle de nos méthodes produit un meilleur résultat. L'on retrouve aussi le classement des méthodes : une combinaison avec la méthode hiérarchique produit de meilleurs résultats que la méthode plate, ce qui semble logique puisque la méthode hiérarchique est intrinsèquement meilleure, 200 classes étant nécessaires pour la méthode plate pour rivaliser avec la méthode hiérarchique utilisant 20 classes.

Si l'amélioration des résultats est en soit une première source de satisfaction, le degré de qualité des résultats obtenus en est une deuxième. En effet cette qualité (F1=84, 10 meilleurs candidats) permet d'envisager dans un très proche avenir l'utilisation concrète de corpus comparables pour l'extraction bilingue, puisque les résultats permettent une révision manuelle similaire à celle nécessaire lors de l'utilisation de corpus parallèles.

Méthode standard combinée avec :	**10 meilleurs**
Plate 1	79
Plate 100	80
Plate 200	83
HIER 10	82
HIER 20	**84**
HIER 50	**84**

Tableau 9 : Corpus médical : évaluation de la combinaison de la méthode standard et des nouvelles méthodes proposées. Un gain de 20% en absolu est alors obtenu.

7. Conclusion

Dans cet article, après avoir opéré un retour sur les hypothèses sous-jacentes aux travaux sur l'extraction de lexiques bilingues de corpus comparables, nous avons proposé une nouvelle méthode qui présente plusieurs avantages par rapport aux méthodes utilisées jusqu'à présent. En particulier, nous avons montré que notre méthode produisait des résultats soit équivalents soit meilleurs que la méthode standard selon le corpus utilisé, et en particulier selon sa taille. De plus, nous avons montré comment une combinaison des deux approches permet d'atteindre des résultats de 20% supérieurs en absolu à chacune des deux méthodes, fournissant des lexiques dont la qualité nous semble permettre la révision par un lexicographe.

Références

[BIBER 93] BIBER D. *Using Registered-diversified Corpora for General Language Studies*, Computational Linguistics,19(2), 1993.

[BLANK 00] BLANK I., *Terminoligy Extraction from Parallel techincal texts*, In Jean Véronis (Éd.), Parallel Text Processing- Alignement and Use of Translation Corpora, Kluwer Academic Publishers, 2000.

[CHUQUET 89] CHUQUET H., Paillard M. *Approche linguistique des problèmes de traduction*. Ophrys, 1989.

[DEERWESTER 90] DEERWESTER S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. *Indexing by latent semantic analysis*. Journal of the American Society for Infomation Science. 1990.

[DEJEAN 02] DEJEAN Hervé, Gaussier Eric, Sadat Fatia, *An approach based on multilingual thesauri and model combination for bilingual lexicon extraction*, International Conference on Computational Linguistics, COLING, 2002.

[DUNNING 93] DUNNIG T, *Accurate Methods for the Statistics of Surprise and Coincidence*. Computational Linguistics,19(1), 1993.

[FABRE 98] FABRE Cécile, HABERT Benoît. *Acquisition de relations entre mots pour une lecture sémantique de corpus*. In: 4èmes journées internationales d'analyse statistique des données textuelles (JADT'98), éd. par Mellet (Sylvie), pp. 273-282. 1998.

[FUNG 98] FUNG Pascale, *A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-Parallel Corpora*, conference of the association of Translation in the americas (AMTA), pages 1-17. 1988.

[FUNG 00] FUNG Pascale, *A Statistical View on Bilingual Lexicon Extraction- From Parallel Corpora to non-parallel corpora*, , In Jean Véronis (Éd.), Parallel Text Processing- Alignement and Use of Translation Corpora, Kluwer Academic Publishers, 2000.

[GAUSSIER 00], GAUSSIER Éric, Hull David, Ait-Moktar Salah, *Term Alignment in use : Machine-aided Human Translation.* , In Jean Véronis (Éd.), Parallel Text Processing- Alignement and Use of Translation Corpora, Kluwer Academic Publishers, 2000.

[GREFENSTETTE 94] GREFENSTETTE Gregory, *Exploration in Automatic Thesaurus Discovery*, Dordrecht :kluwer, 1994.

[GROSS 93] GROSS Gaston. *Classes d'objets et traitement automatique*. In Actes du colloque Informatique et Langue Naturelle ILN'93. 1993.

[HABERT 97] HABERT Benoît, Nazarenko Adeline, Salem André. *Les linguistiques de corpus*. Armand Colin. 1997.

[HARRIS 54] HARRIS Zellig S., Distributional Structure, Word, 10(2-3) :146-162, Traduction française Langage (20), 1970.

[HEID 99] HEID U, *A Linguisti Bootstrapping Approach of Term candidates from German Text*, Terminology,5(2), 1999.

[HIEMSTRA 96] HIEMSTRA D., *Using Statistical Methods to Create a bilingual dictionnary*, Master's thesis, University of Twente. 1996.

[HULL 97] HULL David, *Automatic Identification of Bilingual Terminology Lexicon*, Terminology, 5(2), 1997.

[JELINEK 98] JELINEK F., *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[KILGARIFF 01] KILGARIFF Adam, *Comparing Corpora*, International Journal of Corpus Linguistics, 2001.

[LEWIS 67] LEWIS P.A.W., Baxendale P.B., Bennet J.L. *Statistical discrimination of the synonym/antonym relationship between words*. Journal ofthe ACM, 14-20. 1967.

[MASUICHI 00] MASUICHI R., Flournoy S, Kaufmann S, Peters S., *A bootstrapping method for extracting bilingual text pairs*, International Conference on Computational Linguistics, COLING, 2000.

[MAYNARD 98] MAYNARD D., Ananiadou S. Term Sense Disambiguation Using a Domain-Specific Thesaurus. In Proceedings of 1st Internationa Conference on Language Resources and Evaluation (LREC), Granada, Spain, 1998.

[NEDELLEC 99] NEDELLEC Claire. *Corpus-based learning of semantic relations by the ILP system Asium*. In James Cussens, editor, Learning Language in Logic workshop co-located with ICML'99, pages 28-39, 1999.

[PETERS 95] PETERS C., Picchi E., *Capturing the Comparable : A System for Querying Comparable Text Corpora*, In JADT'95, 1995.

[RAJMAN 00] RAJMAN Martin, BESANCON Romaric, CHAPPELIER Jean-Cédric, *Le modèle DSIR : une approche à base de sémantique distributionnelle pour la recherche documentaire*. Traitement Automatique des Langues, volume 41, numéro 2. 2000.

[RAPP 99] RAPP Reinhard, *Automatic Identification of Word Translations from Unrelated Corpora*, In Proceedings of ACL, 1999.

[SHAHZAD 99] SHAHZAD I., Ohtake K., Masuyama S., Yamamoto K., *Identifying translations of compound nouns using non-aligned Corpora*, MAL'99, 1999.

[TANAKA 96] TANAKA K., Iwasaki H., *Extraction of Lexical Translations from Non-aligned Corpora*, International Conference on Computational Linguistics, COLING, 1996.

[VIVALDI 01], VIVALDI J, RODRIGUEZ H., *Improving Term Extraction by Combining Different Techniques*, Terminology, 7(1), 2001.

[VERONIS 00], VERONIS Jean, Langlais Philippe, *Evaluation of parallel text alignment systems - The ARCADE project. ,* In Jean Véronis (Éd.), Parallel Text Processing- Alignement and Use of Translation Corpora, Kluwer Academic Publishers, 2000.

[WONG 85] WONG S.K., Ziarko W., Wong P.C.N. Generalized Vector Space Model in Information Retrieval. Proceedings of the ACM SIGIR conference, 1985.